Powered by

Futurense

AIL1020

Foundations of Statistics & Probability

Instructor

Dr. Rajlaxmi Chouhan

Associate Professor, Department of Electrical Engineering
Head, Center for Education & Technology for Education
IIT Jodhpur

✉ rajlaxmichouhan@iitj.ac.in

Powered by

Futurense

Module 02

Measures of Central Tendency Contd.

Normal and paired datasets

Recap

Descriptive Statistics

Chebychev’s Inequality

In this video,

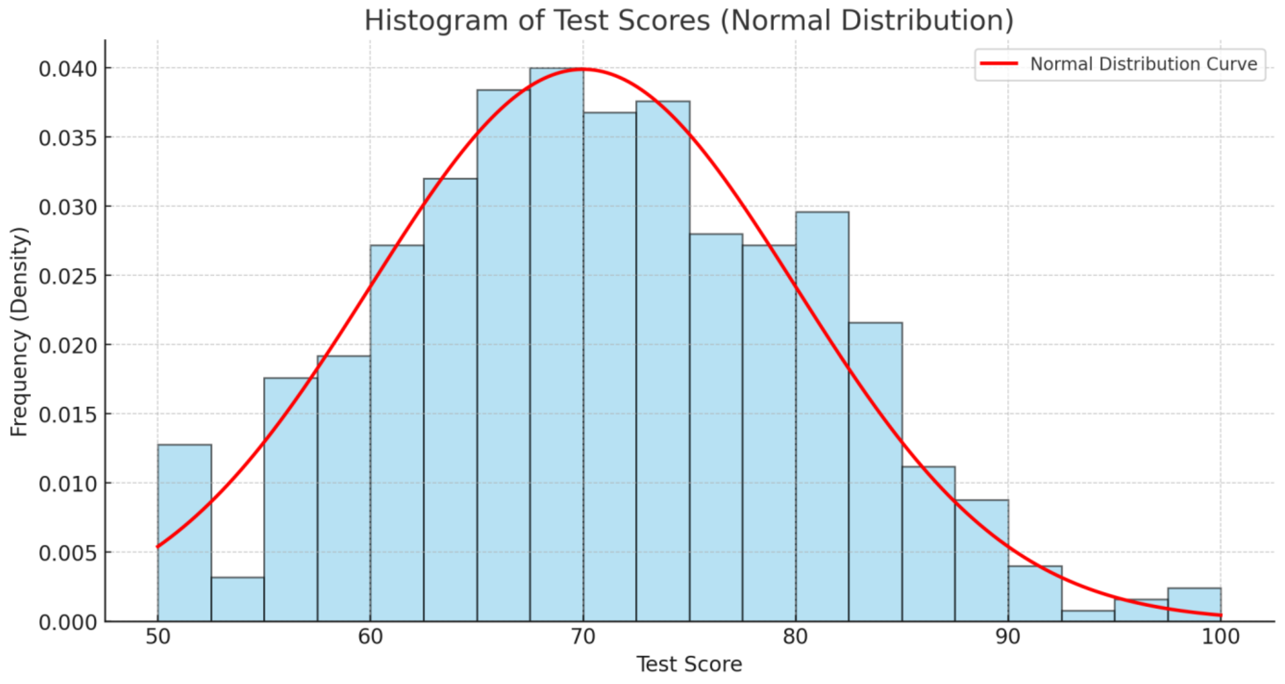
Normal Distribution of Dataset

Paired dataset

Normal Distribution (dataset)

The **normal distribution** describes how data points in a dataset tend to cluster around a central value, with most data points falling close to the mean and fewer occurring as you move further away.

This results in the characteristic **bell-shaped curve** when the data is plotted as a histogram.

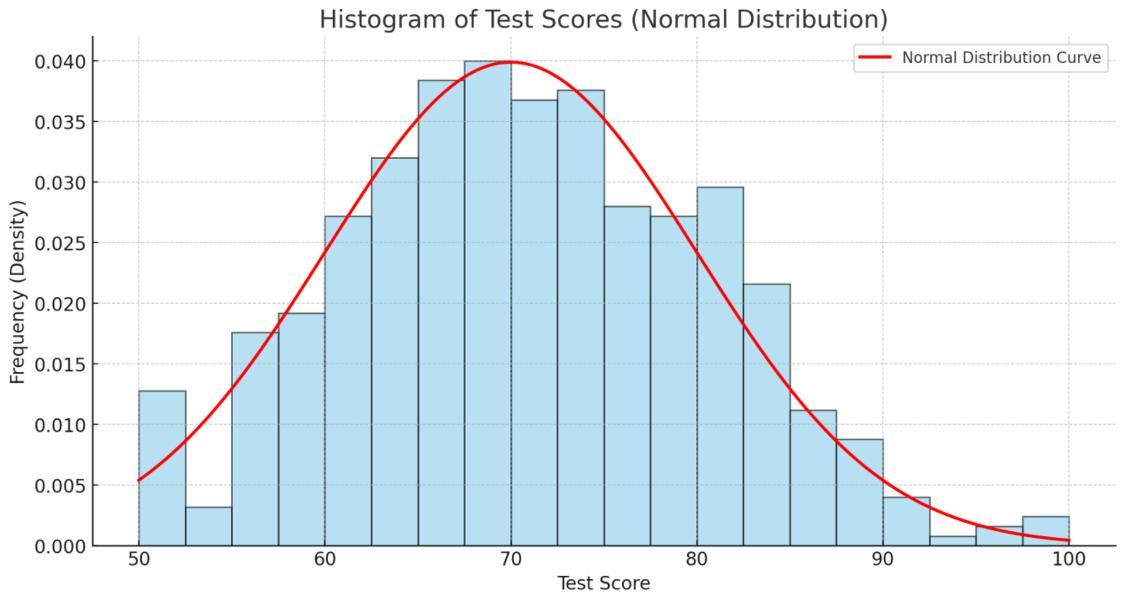


Key Properties of a Normally distributed Dataset

Symmetry: The dataset is evenly distributed around the mean.

Mean, Median, and Mode coincide.

Most data points fall within a predictable range.



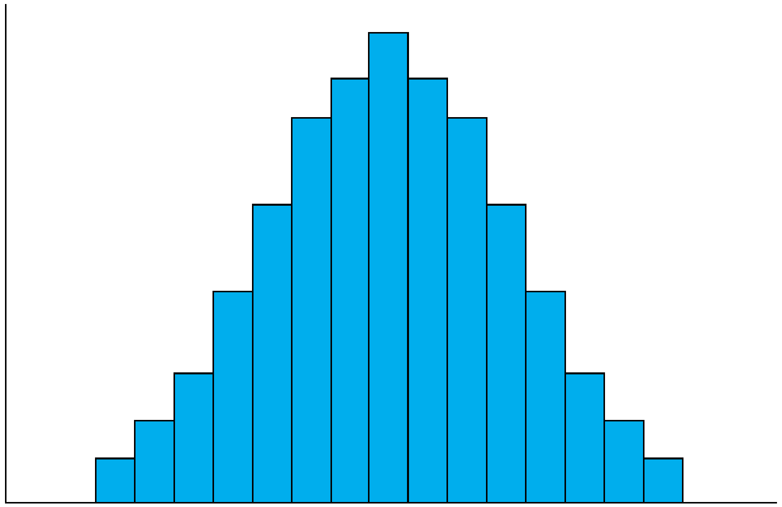
Visualization

A histogram of a dataset that follows a normal distribution shows a peak at the mean and gradually tapers off on both sides:

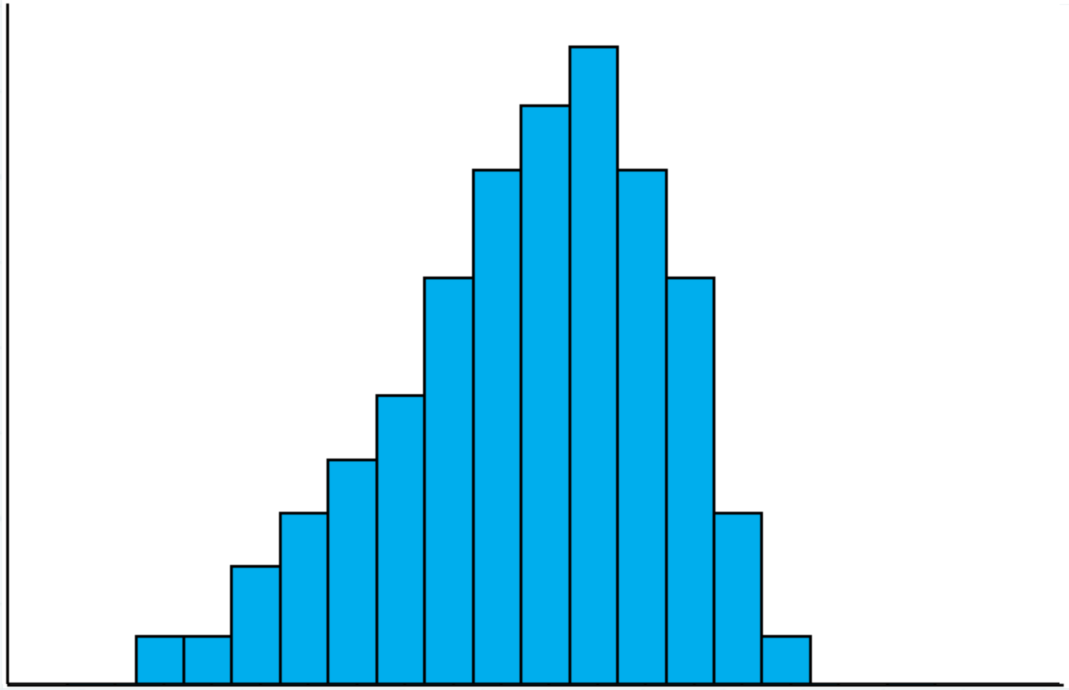
68% of the data falls within one standard deviation (σ) of the mean.

95% falls within two standard deviations.

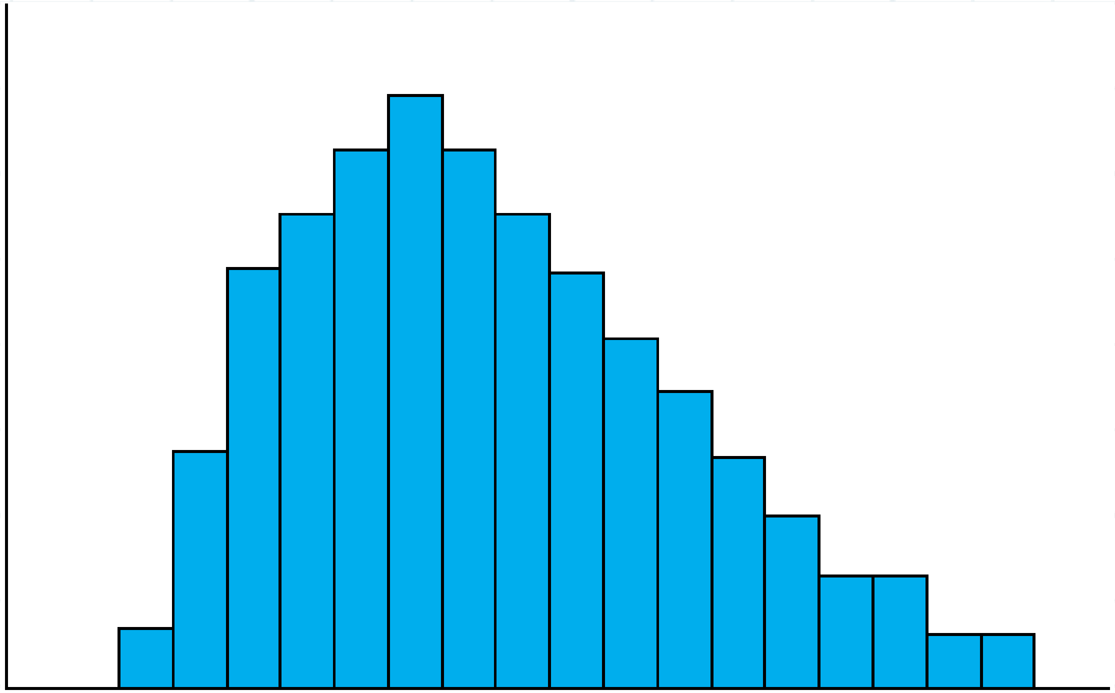
99.7% falls within three standard deviations.



Histogram of Dataset “Skewed”

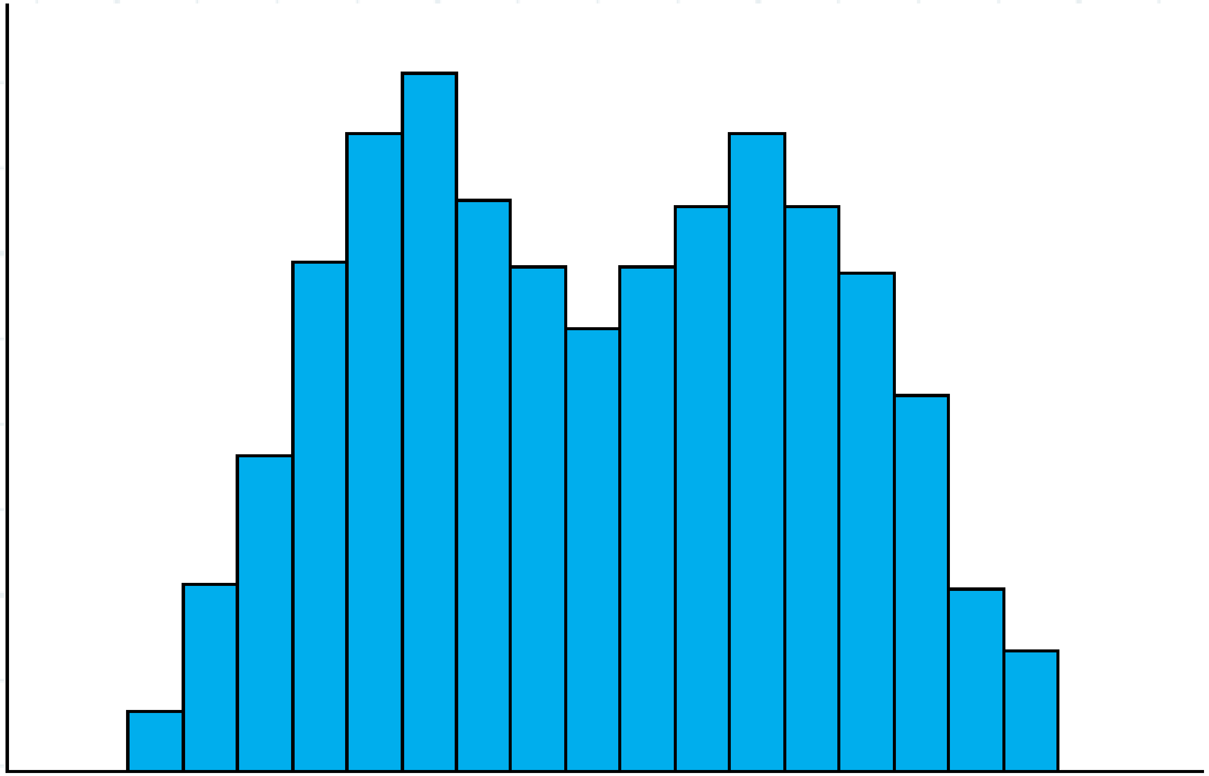


Histogram of dataset skewed to the left



Histogram of dataset skewed to the right

Bimodal Histogram



A histogram with two local peaks: *Bimodal* Dataset

Normal Distribution in Real-world Scenarios

Natural Phenomenon

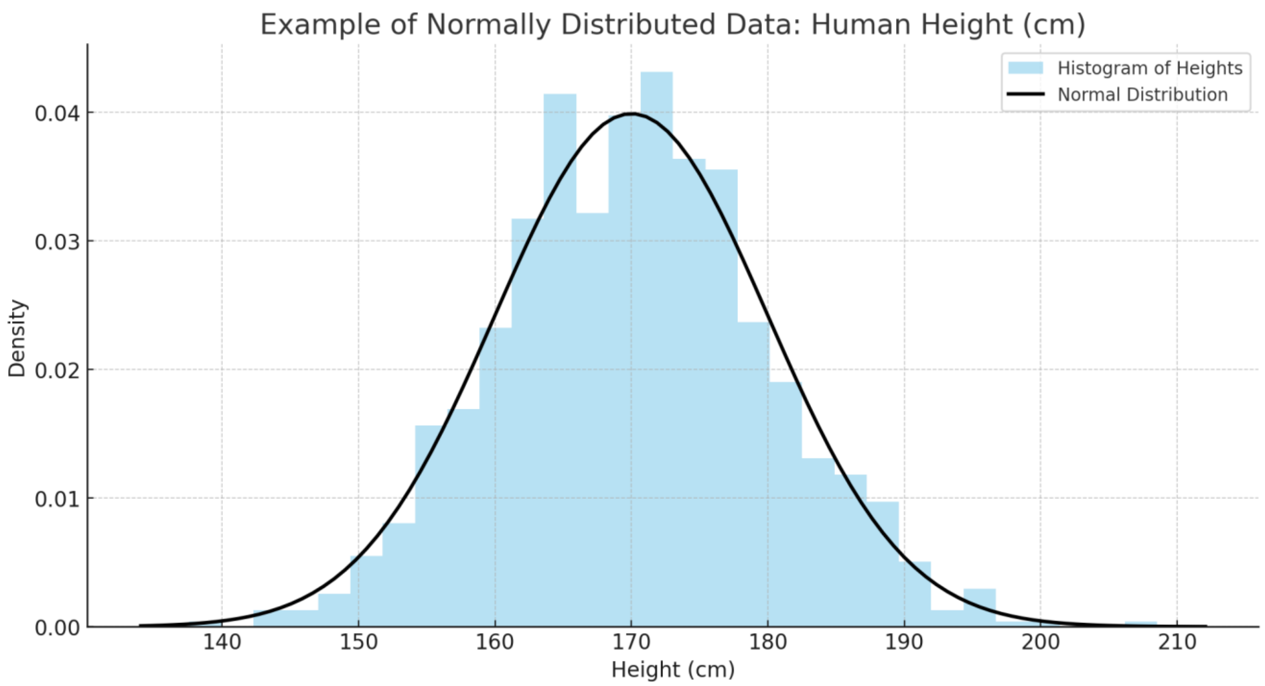
Measurement errors

Social and economic phenomenon

Biological processes

Industrial control processes

Financial data (in some cases)



Central Limit Theorem

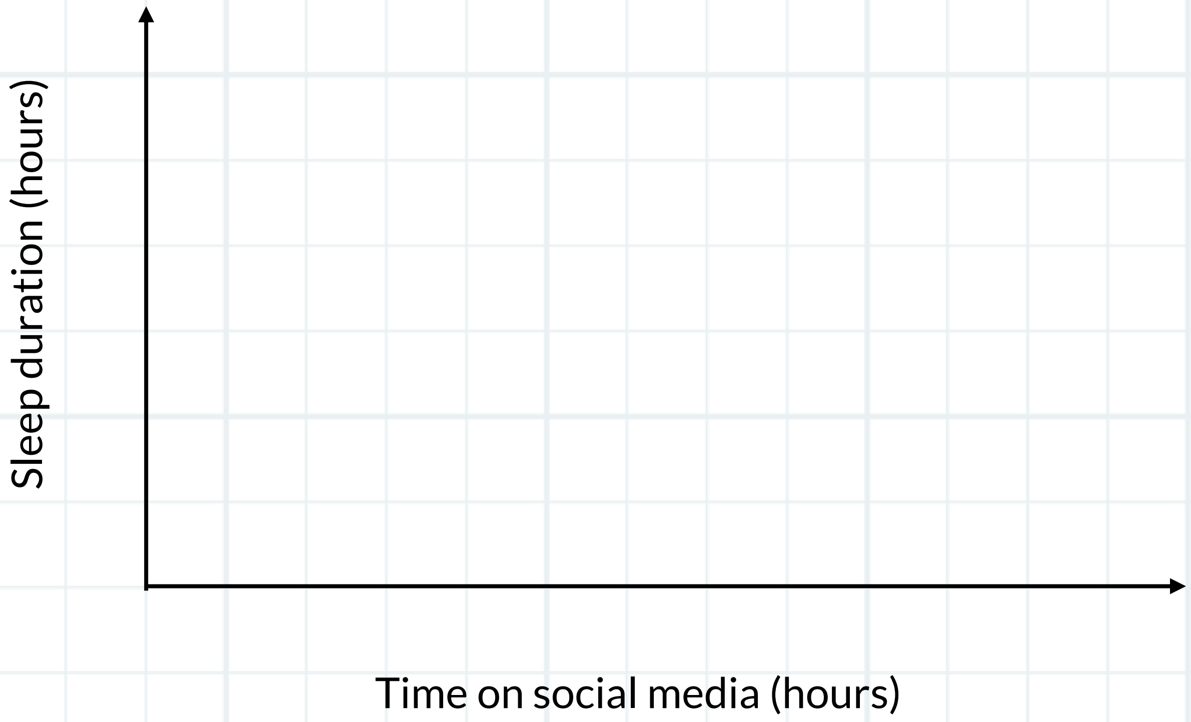
The Central Limit Theorem states that, given a sufficiently **large sample size**, the distribution of the sample mean (or sum) of a random variable will approach a **normal distribution (bell curve)**, **regardless of the original distribution** of the population.

This happens as the sample size *n* increases.

Paired Dataset

A **paired dataset** consists of two related variables collected for each individual or observation.

Paired datasets are best visualized using **Scatter plots**.



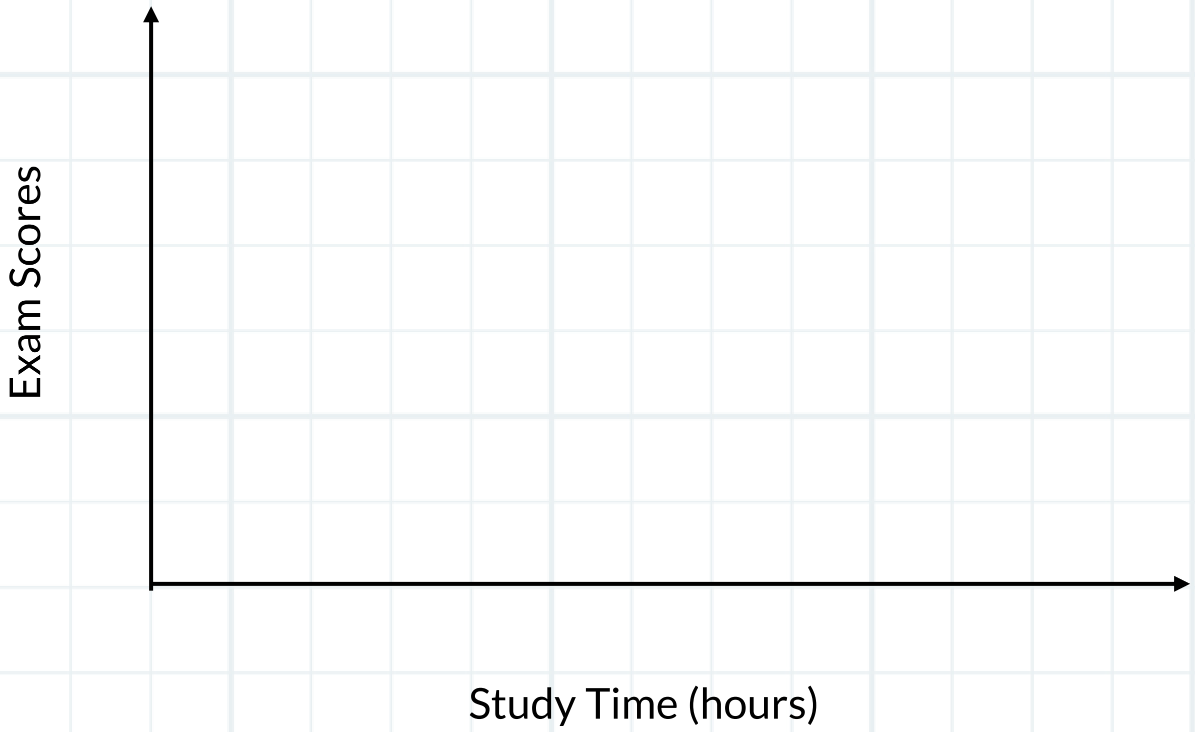
Example of Paired dataset

Friend	Time on Social Media (hours)	Sleep Duration (hours)
1	2	8
2	4	7
3	1	9
4	6	6
5	8	5
6	5	6.5
7	7	5.5
8	3	8
9	9	4.5
10	10	4

Paired Dataset

A **paired dataset** consists of two related variables collected for each individual or observation.

Paired datasets are best visualized using **Scatter plots**.



Another Example of Paired dataset

Friend	Study Time (hours)	Exam Scores
1	8.1	88.1
2	5.4	75.2
3	4.0	68.2
4	2.0	54.9
5	3.1	56.0
6	9.5	94.4
7	3.7	69.7
8	2.5	55.9
9	5.1	75.8
10	3.8	78.3

Scatter Plot

Scatterplots are a great visual tool to identify patterns.

The **shape** of the scatterplot represents the **direction and strength of the relationship** between the two variables.

Scatter Plot

***Slope* of the scatter plot**

A **steeper** slope (either positive or negative) usually indicates a **stronger correlation**.

A **flatter** slope (closer to zero) means a **weaker correlation**.

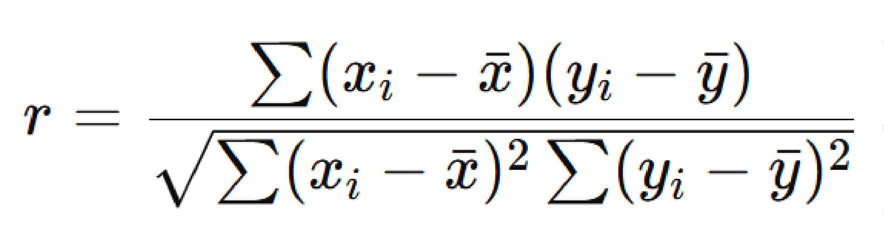
Descriptive Statistics Contd.

Correlation ≠ Causation

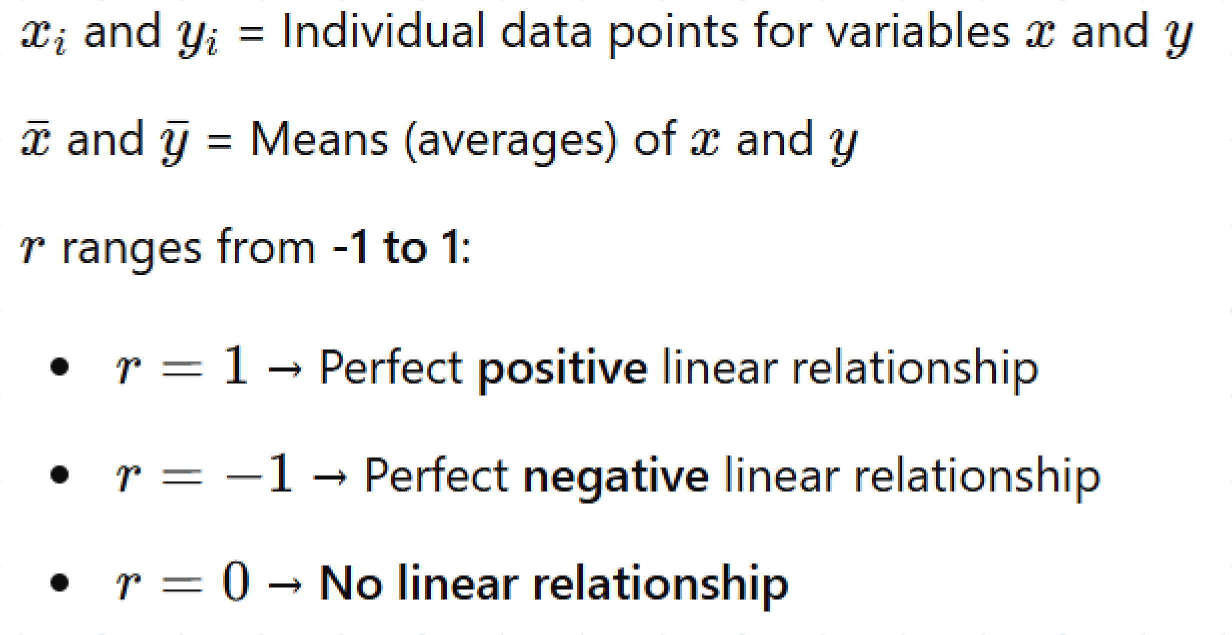
How can **correlation** be quantified?

Sample Correlation Coefficient (*r*)

The **sample correlation coefficient** measures the **strength and direction of a linear relationship** between two variables.



$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



x_i and y_i = Individual data points for variables x and y

$\bar{x}$ and $\bar{y}$ = Means (averages) of x and y

r ranges from **-1 to 1**:

- $r = 1$ → Perfect **positive** linear relationship
- $r = -1$ → Perfect **negative** linear relationship
- $r = 0$ → **No linear relationship**

Recap

Normal distribution

Paired dataset

- Scatterplots and correlation coefficients are great tools for visualizing and quantifying these patterns or relationships.

- Correlation helps identify relationships between variables, but you must always look for possible hidden factors before concluding causation.